Global Predictions and Tests of Erythroid Regulatory Regions

R.C. HARDISON,* F. CHIAROMONTE,* D. KOLBE,* HAO WANG,* H. PETRYKOWSKA,* L. ELNITSKI,* S. YANG,* B. GIARDINE,* Y. ZHANG,* C. RIEMER,* S. SCHWARTZ,* D. HAUSSLER,† K.M. ROSKIN,† R.J. WEBER,† M. DIEKHANS,† W. J. KENT,† M.J. WEISS,‡ J. WELCH,‡ AND W. MILLER*

**Departments of Biochemistry and Molecular Biology, Statistics, Health Evaluation Services, and Computer Science and Engineering, and Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802; † Center for Biomolecular Science and Engineering and Howard Hughes Medical Institute, University of California at Santa Cruz, Santa Cruz, California 95064; and ‡Department of Pediatrics, Children's Hospital of Philadelphia and The University of Pennsylvania, Philadelphia, Pennsylvania 19104*

Determinations of the genomic DNA sequences of human, mouse, and other organisms are landmark achievements, but the major changes in biology and medicine anticipated as a result (Lander 1996) require that a function be assigned to all the important segments within those genomes (Collins et al. 2003). It has long been realized that functional sequences change more slowly than nonfunctional (neutral) DNA sequences over evolutionary time (Kimura 1968; Li et al. 1981). Some gene prediction and assessment algorithms incorporate interspecies sequence alignments into their analysis (see, e.g., Korf et al. 2001; Wiehe et al. 2001; Nekrutenko et al. 2002). This slower rate also can be predictive for sequences involved in gene regulation. One of the early approaches for finding critical sequences within bacteriophage promoters used sequence comparison (Pribnow 1975), and highly conserved noncoding DNA sequences are now commonly used as guides for potential gene regulatory elements (for review, see Hardison 2000; Pennacchio and Rubin 2001).

In this paper, we address two complications to the large-scale application of genomic sequence alignments to predicting *cis*-regulatory modules (CRMs), i.e., discrete sequences such as promoters, enhancers, and silencers that control gene expression. The rate at which neutral DNA changes is highly variable within a genome (Wolfe et al. 1989; Hardison et al. 2003), and thus the amount of change observed needs to be corrected for local variation in the neutral rate. Such a corrected score can be used to compute a probability that a sequence is conserved because of purifying selection (Waterston et al. 2002; Chiaromonte et al., this volume). The second complication is that DNA sequences which do not code for protein (noncoding DNA) can be selected for functions other than a role in regulating gene expression. Examples include genes for noncoding RNAs such as tRNAs and microRNAs. Sequences involved in chromosome dynamics may also be under selection. We describe an approach to find patterns characteristic of gene regulatory sequences within the alignments (Elnitski et al. 2003).

We are applying these analyses of whole-genome sequence alignments to predict regulatory elements of genes expressed during late erythroid differentiation. This is a particularly attractive somatic cell model for mammalian differentiation because morphologically distinct cell types are made during the progress of differentiation and maturation, and several abundant red cell proteins, such as hemoglobins and cytoskeletal proteins, are well-characterized markers of later maturation (Migliaccio and Papayannopoulou 2001). Furthermore, cultured cell lines such as murine erythroleukemia (MEL) cells can be chemically induced to undergo a transition similar to that of proerythroblasts to erythroblasts (Friend et al. 1971). More recently, progenitor cell lines missing a particular transcription factor critical for erythroid differentiation, GATA-1, have been isolated and phenotypically rescued using a conditionally active GATA-1 (Weiss et al. 1997). Thus, we can assay globally for genes responding in these two models for erythroid differentiation, and in the latter case, it is highly likely that early-responding genes are direct targets of GATA-1. We report some initial success applying the computational predictions of CRMs in these somatic cell systems.

ALIGNMENTS OF WHOLE MAMMALIAN GENOMES

The availability of the human (Lander et al. 2001) and mouse (Waterston et al. 2002) genome sequences makes it possible to determine comprehensively which DNA sequences are present in both, which have been inserted or deleted, and which have been altered by nucleotide substitution since primates and rodents diverged. A high-quality assembly of the rat genome sequence is available (International Rat Genome Sequencing Consortium, in prep.), and adding this to the aligned sequences will provide greater resolution on these issues. All the sequences encoding and regulating conserved functions should be found within the sequences common to mouse and human, hence this is the starting point for our search for predicted CRMs.

In our approach to whole-genome alignments, we first find all the meaningful local alignments between the two sequences using the program *blastz*, and then we use *axtBest* to arrange these local alignments into chains that

reflect blocks of conserved synteny, which can be many megabases in length (Schwartz et al. 2003). Further layers of chaining reflect duplications, inversions, and other events (Kent et al. 2003). All sequences in one genome are given the opportunity to align with sequences in the other, hence it is an all-versus-all alignment; no prior deductions about blocks of conserved synteny are used. The scoring parameters have been optimized for long mammalian genomic DNA sequences (Chiaromonte et al. 2002) and can be set to achieve high sensitivity with very little noise (Schwartz et al. 2003). Although it is not possible at present to know definitively that all the homologous sequences have been aligned, it is likely that the vast majority have been.

When interpreting the results of interspecies sequence alignments, it is important to distinguish the part of the genome derived from the last common ancestor, which we refer to as the "ancestral" portion, from the part that arose only along one lineage (Fig. 1). Virtually all lineage-specific insertions result from retrotransposition events (Lander et al. 2001); these are not aligned in our procedure (Schwartz et al. 2003). Lineage-specific segmental duplications comprise about 6% of the human genome (Bailey et al. 2002), but all copies can align with the comparison species. The ancestral portion is the nonrepetitive DNA plus the repeats that were present in the last common ancestor. Orthologous ancestral repeats are included in alignments that begin in adjacent single-copy regions. The ancestral repeats are relics of transposable elements that were active prior to the primate–rodent divergence but no longer transpose. Although a very small fraction of ancestral repeats have been implicated in regulation of gene expression (Jordan et al. 2003), the vast majority have no identifiable function. Hence, the aligned orthologous copies in human and mouse represent a good model for evolution in neutral DNA (Waterston et al. 2002; Hardison et al. 2003).

To the extent that the alignments are comprehensive,

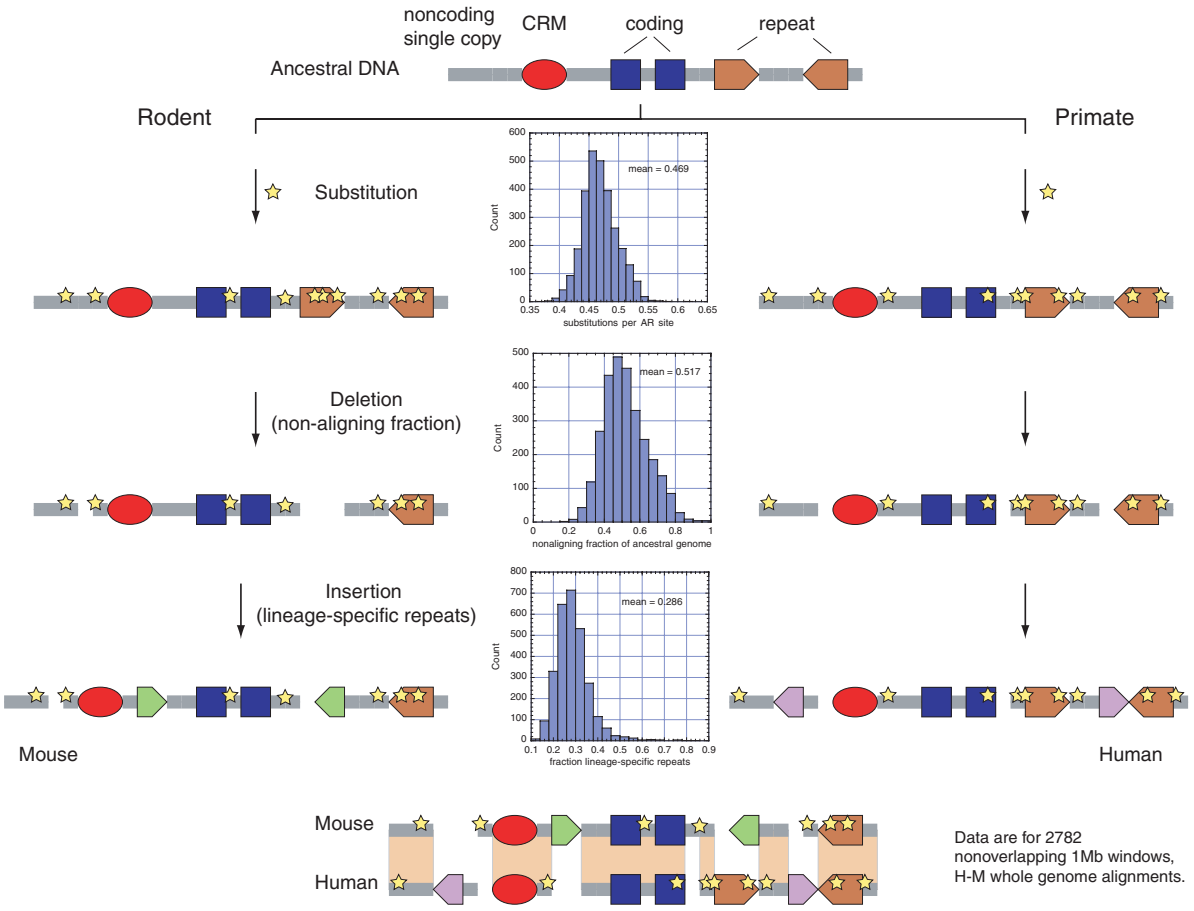

Figure 1. Events that cause sequence divergence over evolutionary time and variation in their rate of occurrence. Types of DNA sequences in the ancestor to rodents and primates are diagramed on the top line; these include coding exons, *cis*-regulatory modules (CRMs), and interspersed repeats from transposons (*brown angled boxes*), separated by single-copy DNA. The diagrams along the sides illustrate the accumulation of nucleotide substitutions (*stars*), large deletions (*absence of icons*), and insertions of new classes of transposons (*green* and *purple angled boxes*) in the lineage to mouse (*left*) and human (*right*). Sequences that were in the ancestor and have not been deleted can align between human and mouse (*bottom diagram*). From the human–mouse alignments, one can infer the substitutions per site in the ancestral repeats (AR), the amount deleted (nonaligning fraction of the ancestral genome), and the amount of DNA inserted (fraction lineage-specific repeats) in 1-Mb nonoverlapping windows (2782 in the human genome). The graphs in the center column show the distribution of these values for the three processes of DNA change.

one can draw an informative inference about the nonaligning part of the ancestral portion of a genome—it is not likely to be present in the other genome. Because we do not align lineage-specific insertions, and lineage-specific duplicates can align, the simplest explanation for the sequences not being in the comparison genome is that they were deleted. Other analyses based on a relatively constant genome size in mammals also argue that the nonaligning fraction reveals deletions in the comparison genome (Waterston et al. 2002).

Genome sequences of additional species, such as rat (International Rat Genome Sequencing Consortium, in prep.), are being assembled as large-scale genomic sequence and analysis projects move into more functional and analytical studies (Collins et al. 2003). Pair-wise and multiple alignments of these sequences are regularly updated and made available on the UCSC Genome Browser (Kent et al. 2002) at http://genome.ucsc.edu. Additional mammalian genome sequences substantially improve the power of sequence alignment techniques to resolve functional from nonfunctional DNA sequences (Thomas et al. 2003).

CONSERVATION AND SELECTION

About 40% of the human genome aligns with sequences in the mouse genome. As expected, almost all (99%) of the genes align between the genomes. These account for at most 2% of the human genome, and they are obviously under selective constraint. The other 38% of the human genome that does not code for protein but still aligns with mouse should include gene regulatory sequences and other functional noncoding sequences. However, these alignments also include much neutral DNA; e.g., about one-fourth of all the ancestral repeats in humans align with orthologs in mouse. All the sequences that align between mouse and human are conserved in the sense that they are present in both species, but the goal is to identify the sequences that are subject to purifying selection. It is the latter sequences that are playing a role in some conserved function.

A major complication to answering this question is that the rate of neutral evolution varies across the genome. The distribution of nucleotide substitutions per site in ancestral repeats (computed on 1-Mb nonoverlapping windows) is quite wide (Fig. 1), reflecting substantial regional variation in the underlying neutral substitution rate. In addition, the amount of DNA inferred to be deleted from mouse and the amount of transposable elements inserted and retained show substantial variation (Fig. 1). Furthermore, the amounts of neutral substitution, deletion, insertion (of LTR repeats), recombination, and single-nucleotide polymorphisms (SNPs) covary dramatically (Fig. 2A). Because the substitutions are measured in neutral DNA, different levels of selection cannot ex-

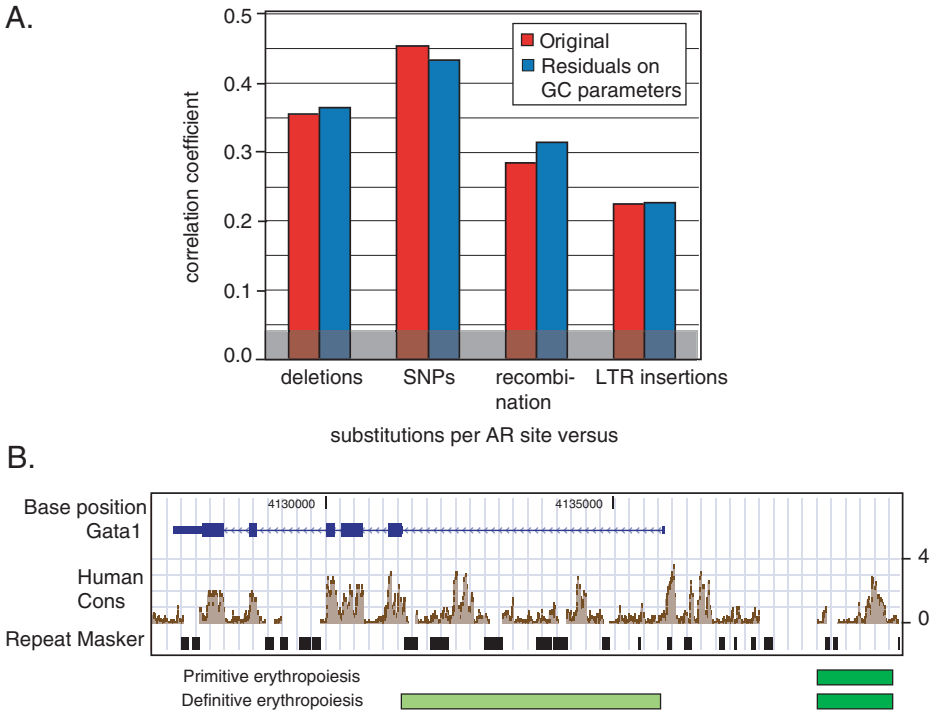

Figure 2. Covariation in divergence rates and application of an alignment score that accounts for local rate variation. (*A*) Correlation among the amounts of neutral substitution (in ARs) with large deletions (based on human–mouse alignments), insertions of LTR repeats, recombination, and SNPs in human. Correlations are shown for the original data and for residuals after the quadratic effects of fraction G+C, change in fraction G+C, and CpG island density have been removed (Hardison et al. 2003). The correlations of various divergence processes with insertion and retention of different classes of repeats is the subject of ongoing work. (*B*) The *Human Cons* track plots the *L* score giving the log-likelihood that alignments reflect selection for the mouse *Gata1* gene. The gene is transcribed from right to left. The upstream and intronic noncoding regions with high *L* scores correspond to previously described strong erythroid enhancers (Onodera et al. 1997).

plain the regional differences. Rather, the covariation in the various divergence processes appears to reflect an inherent tendency of large, megabase-sized regions to change at a fast or slow rate (Chiaromonte et al. 2001). The molecular and cellular basis for this inherent tendency to change is unknown, although it is possible that repair of double-stranded breaks could be at least part of the explanation (Lercher and Hurst 2002).

Given the variation in neutral substitution rates, the goal is to find aligning segments whose similarity significantly exceeds that expected from divergence at the local neutral rate. These should be the sequences subject to purifying selection. Indeed, the significance of a particular alignment score will vary substantially depending on the divergence rate of the surrounding DNA (Li and Miller 2002). Thus, the fraction of matching nucleotides for alignments in small (50 bp) windows was adjusted for the local neutral rate, empirically estimated from nearby aligning ancestral repeats. The overall distribution of these adjusted scores is broad; when compared to its neutral component (the distribution for ancestral repeats only) it presents a marked right-skewedness—i.e., increased frequencies on higher score values (Waterston et al. 2002). A statistical decomposition of this skewed overall distribution leads to the conclusion that about 5% of the human genome is under purifying selection (Waterston et al. 2002; Chiaromonte et al., this volume). This is over twice the amount of DNA that codes for protein, showing that the noncoding portion of the genome contributes significantly to the functional DNA. However, it is only about one-eighth of the conserved sequences, so a majority of the aligning sequences do not reflect selection for some function.

To make these scores more useful to biomedical scientists, *L* scores (or *Mouse Cons* and *Human Cons*) have been computed that convert locally adjusted similarity scores into probabilities that alignments in a given 50 bp result from selection. These can be accessed at the UCSC Genome Browser. An example of this track for the mouse *Gata1* gene shows that protein-coding exons, the first intron, the promoter, and a region about 3–4 kb further upstream are not only conserved (align), but are highly likely to be generated by selection (Fig. 2B). The upstream region corresponds to an enhancer that confers erythroid-specific expression during primitive erythropoiesis and collaborates with the intronic enhancer to activate expression during definitive erythropoiesis (Onodera et al. 1997). Thus, these scores, generated from the alignment scores adjusted for local rate variation, can be effective indicators of CRMs.

DISCRIMINATING CRMS FROM OTHER DNAs

The *Mouse/Human Cons* or *L* scores are measures of alignment quality, where matches are favored more than mismatches, which are favored more than gaps. Noncoding DNA sequences with a high *L* score are more likely to be subject to purifying selection, and this set of sequences should contain CRMs regulating conserved functions. However, it should also contain other functional sequences such as genes encoding structural RNAs and microRNAs.

Therefore, we explored several computational approaches to analyzing interspecies genomic sequence alignments, aiming to develop computational methods to distinguish regulatory regions from neutrally evolving DNA. To do so, we employed statistical models that recognize alignment patterns characteristic of those seen in known CRMs. Alignments rich in these patterns need not be those that score highest in quality (e.g., a similarity score) or a likelihood of being under selection. Known enhancers and other CRMs tend to be clusters of highly conserved binding sites for transcription factors, but sequences between those binding sites are more variable between species. Thus, alignment quality measurements in the CRMs are usually less than those seen in regions under more uniform selection, such as coding exons.

Three training sets were collected from the whole-genome human–mouse alignments: (1) known CRMs, which are a set of 93 experimentally defined mammalian gene regulatory regions (accessible from GALA at http://www.bx.psu.edu/), (2) well-characterized exons (coding sequences, as a positive control), and (3) ancestral interspersed repeats (the major sequence class used for neutrally evolving DNA). Quantitative evaluation of statistical models that potentially could distinguish functional noncoding sequences from neutral DNA showed that discrimination based on frequencies of individual nucleotide pairs or gaps (i.e., of possible alignment columns) is only partially successful. In contrast, scoring procedures that include the alignment context, based on frequencies of short runs of alignment columns, achieve good separation between regulatory and neutral features (Elnitski et al. 2003).

The best-performing scoring function, called regulatory potential (*RP*) score, employs transition probabilities from two Markov models estimated on the training data. In practice, the procedure evaluates short strings of columns in the alignments, giving a higher value to those that occur more frequently in the CRMs training set than in the ancestral repeats set (Fig. 3). In this procedure, alignments are described using a reduced alphabet *A*. In each training set, we compute the frequencies with which short strings of alignment characters are followed by a particular alignment character. As an example, consider alignment columns to consist of two types of matches, those that involve G or C (S) and those that involve A or T (W), plus transitions (I), transversions (V), and gaps (G). A 5-symbol alphabet can thus describe the alignments. For short strings, the number of possible arrangements of these 5 symbols is computationally manageable. Therefore, we estimate the probability that any string of length *T* is followed by a particular symbol (transition probabilities), where *T* is the order of the Markov model. For example, the empirical frequencies of a pentamer, say WIISV, followed by a given symbol, say S (or W, I, V, G) are used to estimate the transition probabilities of a fifth-order Markov model.

More generally, for a Markov model of order *T*, we estimate the probability that within a regulatory region an

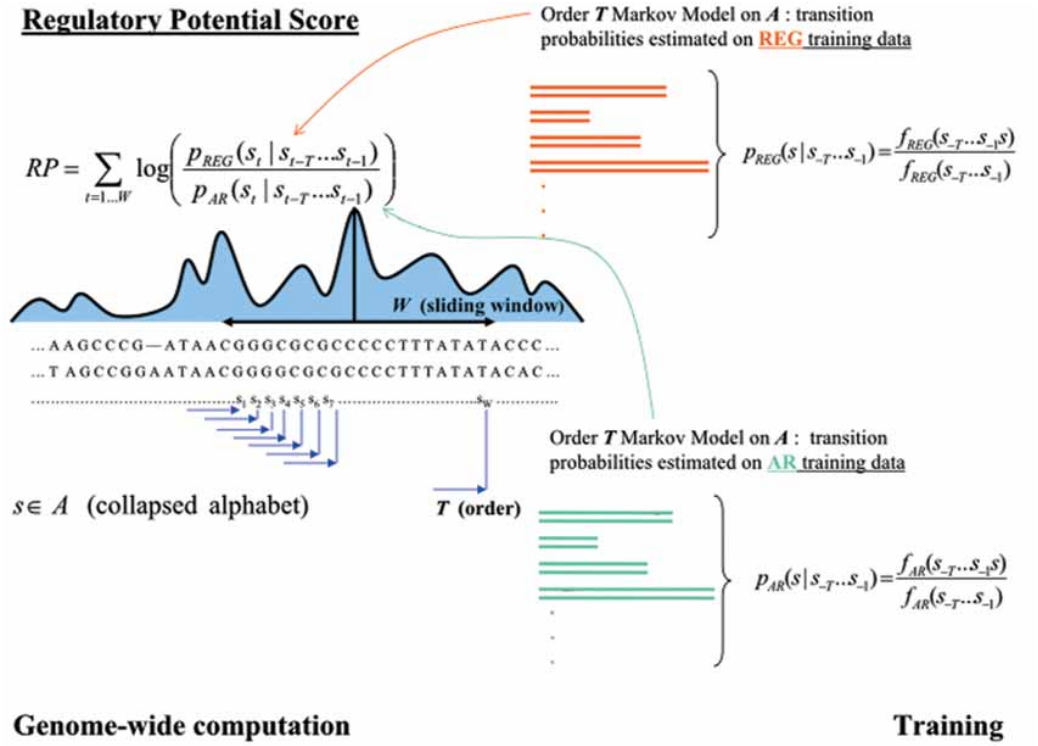

Figure 3. Genome-wide computation of regulatory potential (*RP*) scores. Diagrams on the right illustrate the use of two training sets, known regulatory regions (REG) and ancestral repeats (AR), to build Markov models describing the likelihood of a string of *T* alignment characters being followed by a particular alignment character. The alignment characters are from a collapsed alphabet (*A*) that describes mismatches, gaps, and different kinds of matches. The diagram on the left illustrates the application of these Markov models to calculate the log-likelihood that a segment of an alignment (window size *W*) fits with the model for a regulatory region rather than an ancestral repeat. This log-likelihood is the regulatory potential.

alignment character s is preceded by the string of characters s_{-T} to s_{-1} ($p_{REG}[s/s_{-T}...s_{-1}$] in Fig. 3) as the empirically observed frequency of the string $s_{-T}...s_{-1}s$ divided by that of the string comprising the first *T* positions, $s_{-T}...s_{-1}$, in the CRMs training set. We then repeat the same estimation procedure on the ancestral repeats (*AR*) training set.

The *RP* score is computed for any alignment by dividing the transition probability for regulatory regions, $p_{REG}(s/s_{-T}...s_{-1})$, by that for ancestral repeats, $p_{AR}(s/s_{-T}...s_{-1})$, at each position in the alignment, taking the logarithm, and summing over positions. This log-odds ratio is illustrated in Figure 3 for sliding windows of length *W*. When needed, the score is adjusted for the length of the alignment (Elnitski et al. 2003). The *RP* score has been computed in 50-bp windows (overlapping by 45 bp) for the human–mouse whole-genome alignments, using a 5-symbol collapsed alphabet and a fifth-order Markov model. These scores and plots of them are provided at the UCSC Genome Browser (http://genome.ucsc.edu, Nov. 2002 human assembly), and they are recorded in the database of *g*enomic DNA sequence *al*ignments and *a*nnotations, GALA (Giardine et al. 2003).

A validation study shows that this approach can separate the reference data set of 93 known regulatory regions from the ancestral repeat segments used in training (Fig. 4). Cross-validation studies also support the discriminatory power of the *RP* score (Elnitski et al. 2003). Of note, the accuracy of our predictive models should become even greater as additional regulatory sequences demonstrated through experimental approaches are added to the training set and as more alignments are added. Moreover, the same computational approach can be applied to discrimination among other functional classes, as training data from them become available.

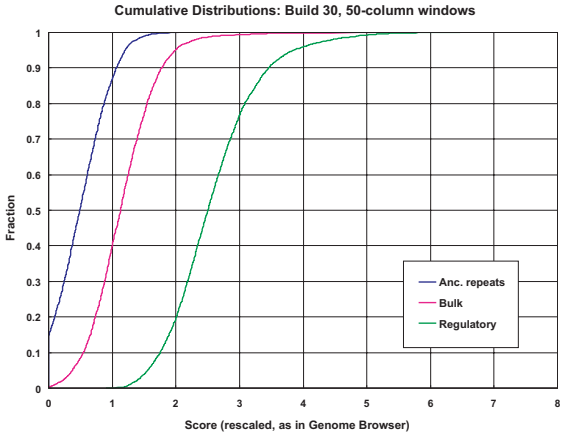

Figure 4. Cumulative distribution of *RP* scores of alignments in several classes of DNA, evaluated using fifth-order Markov models and a 5-letter alphabet. Note the complete separation between regulatory regions and neutral DNA (ancestral repeats). The "bulk" alignments are a set picked at random from all alignments.

CALIBRATION OF THE REGULATORY POTENTIAL SCORE

Realizing that performance on the training set is seldom indicative of performance on new problems not already in the training set, we analyzed the ability of the *RP* score to find known regulatory regions in a well-studied gene complex. The goal is to find an optimal threshold for the *RP* score such that known CRMs are found with high efficiency (high sensitivity) while other noncoding sequences are largely excluded (high specificity). The complex of β-like globin genes (the *HBB* complex) on human Chromosome 11 was chosen for these calibration studies because proximal promoters and upstream regulatory sequences (within a few hundred base pairs of the promoters) have been identified for each active gene, and high-level expression of all the genes is dependent on a distal (as much as 60 kb upstream) enhancer called the locus control region, or LCR (for review, see Forget 2001; Hardison 2001; Stamatoyannopoulos 2001). The LCR is marked by at least four DNase hypersensitive sites (HS1–HS4) that contribute individually and collectively to enhancer function (for review, see Hardison et al. 1997; Li et al. 2002). The five active genes are transcribed right to left in the diagram in Figure 5. A set of eleven intervals was compiled that cover each of the well-characterized CRMs for which experiments show clear, independent effects on regulation. DNA sequences that affect expression levels only in combination with other CRMs were not included. Four of the eleven intervals in the reference set were also in the training set used for the *RP* score. This limits the stringency of this test, but until a larger number of regulatory regions are carefully characterized, some overlap with the training set is difficult to avoid. The reference CRMs are covered by pair-wise and 3-way alignment scores and by the *RP* score, but with different values. Some CRMs, such as the LCR HS3 and the upstream regulatory regions of *HBBG1* and *HBBG2*, have higher *RP* scores than conservation scores.

We used the GALA database (Giardine et al. 2003) to organize and extract the necessary information for the calibration study to find an optimal *RP* threshold. GALA is a relational database with genome-wide information on genes (known and predicted), exons, gene products (including Gene Ontology descriptions; Ashburner et al. 2000), gene expression (including the GNF data using Affymetrix human and mouse gene chips; Su et al. 2002), human–mouse alignments, scores such as *L* and *RP* derived from the alignments, binding sites for transcription factors predicted by matches to *TRANSFAC* weight matrices (Matys et al. 2003), repeats (Smit and Green 1999), and much other information. All data are organized by sequence positions in the human or mouse genome assemblies. GALA allows queries across fields and supports complex queries that combine results from simple queries by conventional set operations (union, intersection, and subtraction) as well as by proximity and by clustering. Thus, it greatly expands the data-mining capacity beyond the conventional one-gene or one-locus view most commonly used at genome browsers. It can be accessed at http://www.bx.psu.edu/.

To determine the *RP* score threshold that works best in identifying the reference set, we queried GALA to find all the ranges of DNA that pass each candidate *RP* score

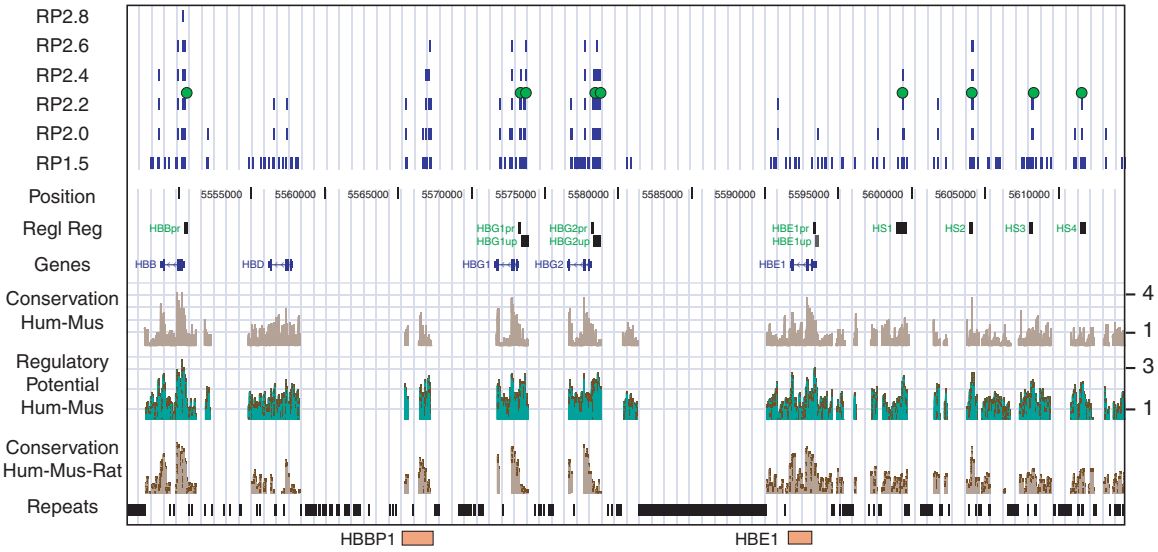

Figure 5. Effectiveness of different regulatory potential (*RP*) thresholds in predicting known regulatory elements in the *HBB* complex. Genes are transcribed from right to left. Conservation scores and *RP* scores are plotted, and genes, repeats, and known regulatory regions are shown on the lines below the position track. Above it are the segments whose *RP* scores exceed the designated threshold (after subtracting exons). Small green circles mark the "true positives" for the RP2.2 track. The failure to find the *HBE1* promoter and upstream regulatory region is an artifact of the current genome annotation, and we have installed a work-around in GALA for future analysis. The results are displayed from the UCSC Genome Browser; comparable analyses are available genome-wide. The results for each *RP* threshold were displayed at the Genome Browser, saved as pdfs, and then combined using Adobe Illustrator. The "Conservation Hum-Mus-Rat" track quantifies the level of conservation in human–mouse–rat alignments (M. Blanchette et al., in prep.).

threshold in the 68-kb interval encompassing the *HBB* complex, including the LCR. After subtracting the exons, the set of DNA intervals passing the threshold were diagramed using an automatic connection between GALA and the UCSC Genome Browser (Fig. 5). As expected, higher thresholds returned fewer intervals, and these sets were enriched in the reference CRMs. For *RP* = 2.2, nine of the eleven reference CRMs are returned. The two that are missing (promoter and upstream regulatory region of *HBE1*) are artifacts of the annotation. They were lost because the annotation of this gene uses a minor promoter in the upstream region, thereby including the CRMs in the annotated "first exon." The "false positives," i.e., intervals meeting the filtering thresholds but not annotated as regulatory regions, are a mixture of overprediction, true regulatory regions that have not been tested, and a few artifacts of incomplete annotation, such as the pseudogene *HBBP1*, which is not in the annotation but whose exons pass the filters.

Detailed comparison between the intervals passing the filters and the reference CRMs shows that the specificity reaches a plateau around *RP* = 2.3 whereas sensitivity declines above this threshold (Fig. 6, left). Indeed, *RP* = 2.3 is a minimum in a cost function (Fig. 6, right), and hence we have used it as the threshold in further analysis. Further analysis using the clustering and proximity capabilities in GALA showed that combining this optimal *RP* threshold with a requirement that a DNA segment have a predicted binding site for GATA-1 improved the specificity from about 0.6 to 0.7. The upper limit on specificity is caused partly by incomplete analysis of potential *cis*-regulatory elements even in the *HBB* complex.

EXPERIMENTAL TESTS OF PREDICTED *cis*-REGULATORY MODULES

The publicly available *RP*, *L*, and other scores can be combined with predictions of binding sites for any relevant transcription factors to predict CRMs genome-wide for a wide variety of mammalian tissues or stages of development. We have begun an extensive set of tests of the predicted CRMs for genes induced during late erythroid differentiation and maturation using the two somatic cell models mentioned in the introduction. Extensive analysis of microarray expression data has revealed a cohort of genes induced along with the β-globin genes (*Hbb-b1* and *Hbb-b2*) in both cell lines. Because the G1E cell line is responding directly to restoration of the activity of the GATA-1 transcription factor, we include predicted binding sites for GATA-1 in our predictions of CRMs. The cohort of coexpressed genes includes some previously known to be induced in erythroid cells, such as *Alas2*, which encodes the enzyme catalyzing the rate-limiting step in heme biosynthesis. Other genes such as *Hipk2* were not well known as erythroid-induced genes.

The gene *Alas2* is an example of our early predictions and tests. Using GALA to search for intervals meeting our criteria (*RP*-score at least 2.3, no exons, and a predicted GATA-1 site within 50 bp), we found four regions in the roughly 25-kb region encompassing human *ALAS2* (Fig. 7, top). (GALA for mouse with mouse–human *RP* scores is now available so that one can perform the analysis entirely from the perspective of the mouse genome.) One predicted CRM is the major promoter, another is in intron 8, which others have shown is an enhancer (Surinya et al. 1998). We focused on a predicted CRM in intron 1 for testing. A more detailed view from our interactive alignment viewer *Laj* (Wilson et al. 2001) shows that the region is strongly conserved and has a predicted GATA-1-binding site in both human and mouse (Fig. 7, bottom).

The strategy for testing the predicted CRMs for enhancer and silencer function is to add them to an expression cassette in which the green fluorescent protein gene is transcribed from a minimal *HBB* promoter, and then we force the test construct to integrate at a marked site in

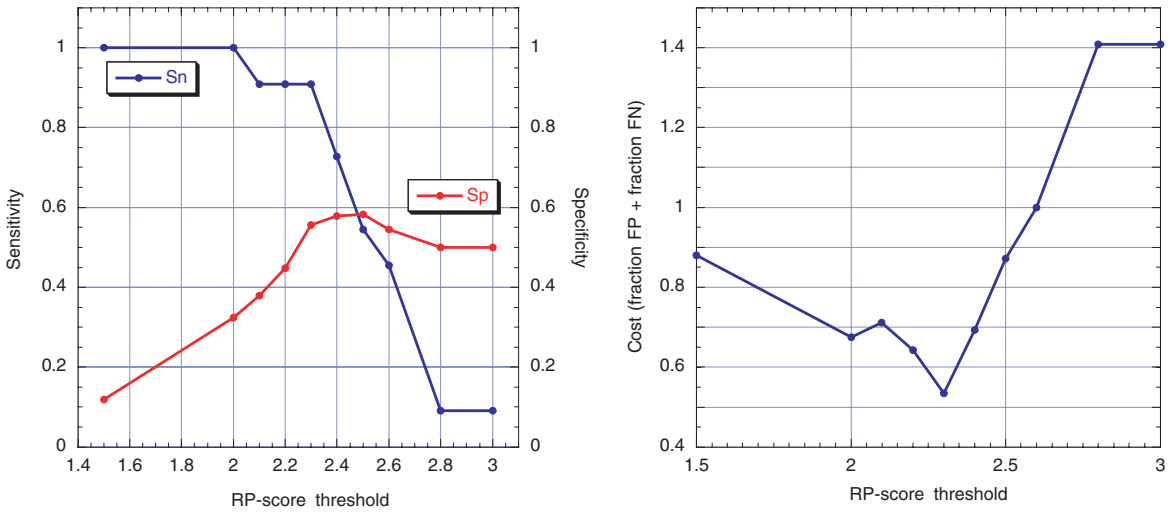

Figure 6. Sensitivity and specificity (*left*) and cost (*right*) of *RP*-score thresholds applied to known regulatory elements in the *HBB* complex. The sensitivity (Sn) is the fraction of known elements found above the indicated threshold, and the specificity is the fraction of segments above the indicated threshold that are known regulatory elements. The cost is the fraction of segments above the indicated threshold that are "false positives" plus the fraction of known elements that are false negatives.

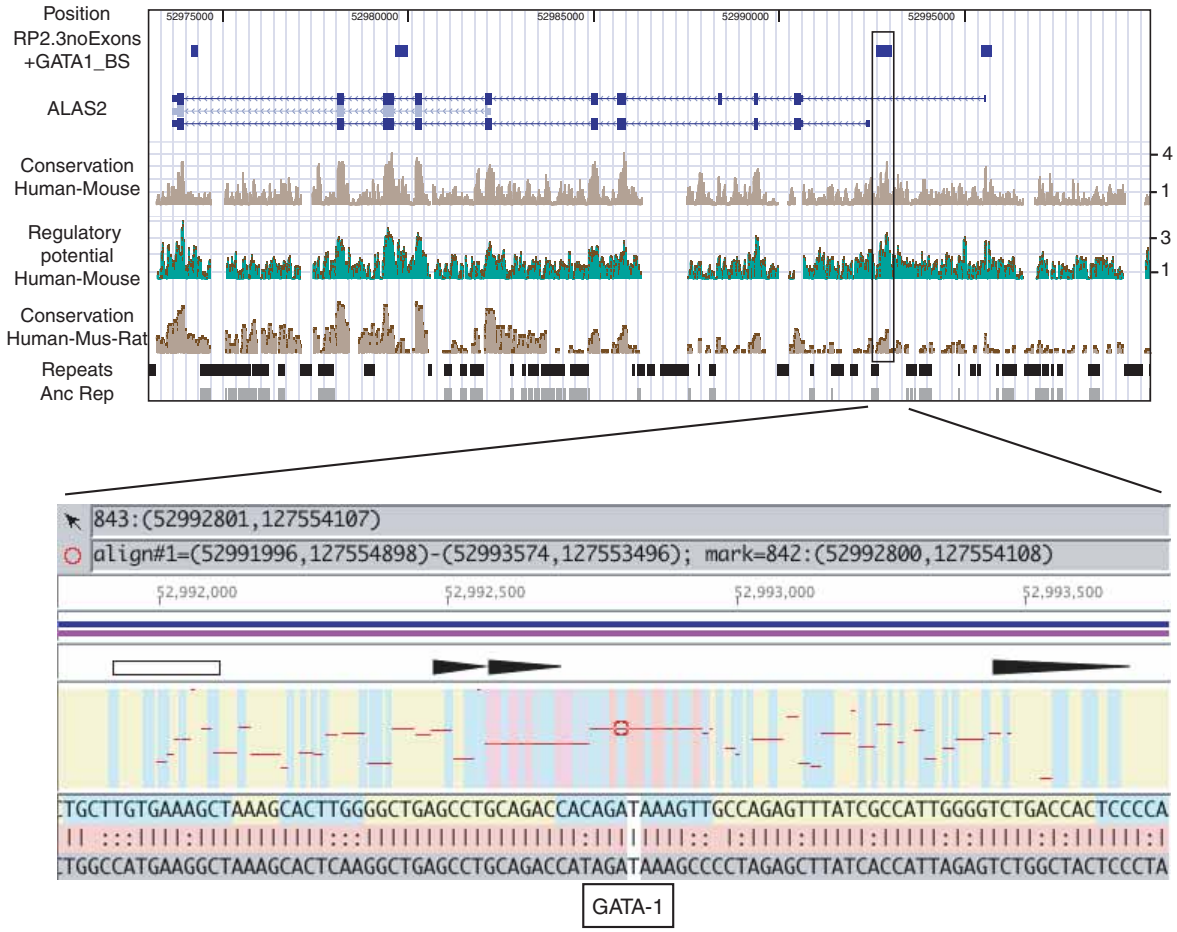

Figure 7. Predicted *cis*-regulatory modules for the *ALAS2* gene (first track under the positions) along with the gene structure (transcribed right to left, exons are boxes) and plots of conservation and regulatory potential (*RP*)-scores (*top*). This output was generated by GALA plus the UCSC Genome Browser. The boxed region was selected for experimental tests. The alignment in this region is shown in the lower panel, which is from our interactive alignment viewer *Laj* (Wilson et al. 2001). Note the presence of a conserved predicted GATA-1-binding site. GALA output can be viewed in the UCSC Browser, *Laj*, or other formats.

MEL cells (Bouhassira et al. 1997; Feng et al. 1999; Molete et al. 2001). Thus, we monitor expression levels in a very precise and accurate way, because the test constructs and parental constructs are at the same chromosomal position (Fig. 8A). The predicted CRM in intron 1 of the mouse ortholog, *Alas2*, was amplified by PCR from murine genomic DNA and ligated into the expression cassette. Both the test and parental expression cassettes were forced to integrate at locus RL5 in MEL cells.

The parental cassette with no enhancer expresses GFP, and it induces a small amount when cells are treated with HMBA (Fig. 8B). Addition of the predicted CRM from *Alas2* intron 1 increased the expression both before induction, showing an enhancer function, and after induction, showing that it also affects inducibility. The effect gets stronger with time of induction.

The predicted CRM is thus experimentally verified as an enhancer and as a sequence that confers erythroid inducibility. At this time, we have tested three predicted CRMs from three different genes and have found that two boost expression and one has no effect in this system. As larger numbers of predicted CRMs are tested, we will be able to improve our models for predicting CRMs.

CONCLUSIONS

Pair-wise and multiple whole-genome alignments of human, mouse, and rat are available and are being continually updated. These alignments, plus scores that provide guidance on the likelihood that a region is under selection or is rich in alignment patterns typical of *cis*-regulatory elements, are available at the Genome Browser and in our GALA database. Users can access these scores alone or in combination with other genomic features of interest (such as predicted binding sites for transcription factors, CpG islands, and exons) to find candidate CRMs throughout the genome. Calibration studies indicate that an *RP* threshold of 2.3 is effective for the *HBB* complex. Our early results on experimental tests based solely on these predictions are encouraging, and we expect continuing improvements in the predictive algorithms. The combination of bioinformatic predictions and

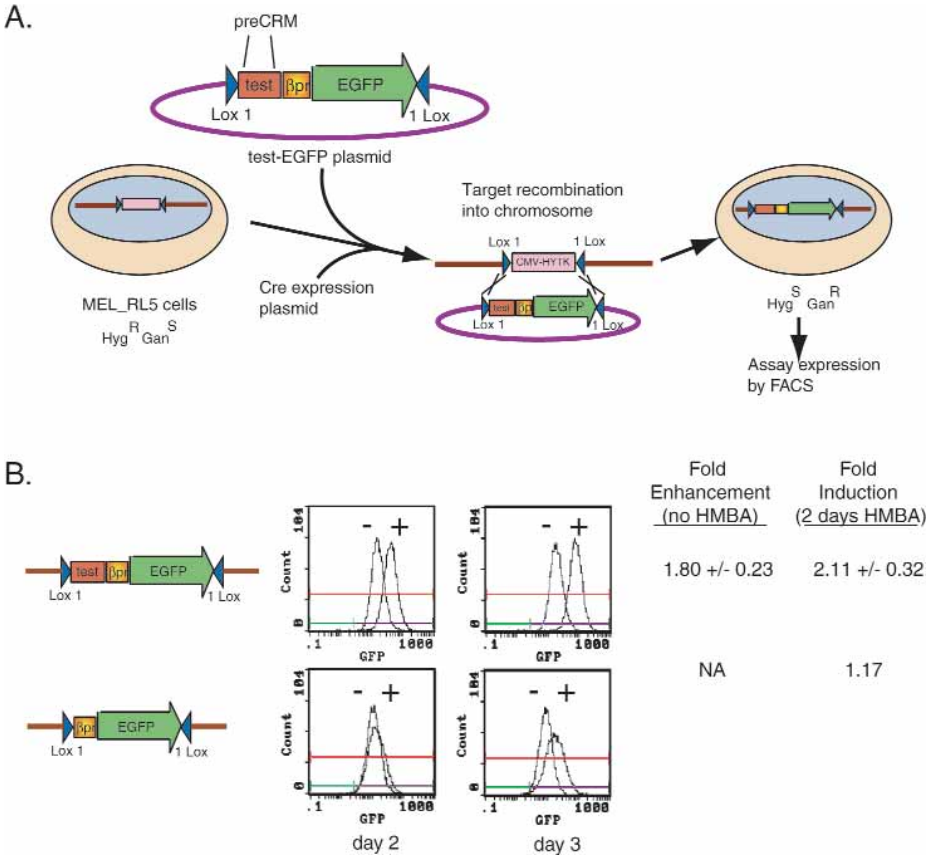

Figure 8. The predicted *cis*-regulatory module in intron 1 of *Alas2* enhances expression and increases inducibility. (*A*) The strategy for isolating and testing predicted CRMs using recombinase-mediated cassette exchange at locus RL5 of MEL cells (Feng et al. 1999) is shown. This procedure ensures that expression of all constructs is monitored after integration at the same chromosomal position. (*B*) Maps of the test and parental expression cassettes are on the left, and the FACS profiles of fluorescence from EGFP are plotted for cells uninduced (–) or induced (+) to erythroid maturation by treatment with HMBA. The fold enhancement and induction are shown on the right.

experimental tests in somatic cell developmental models can serve as a paradigm for global analysis of regulation in any tissue.

ACKNOWLEDGMENTS

We thank the members of all the genome-sequencing consortia for determining the sequences and making them publicly available rapidly. R.C.H., S.Y., F.C., L.E., D.K., S.S., and W.M. were supported by National Human Genome Research Institute grant HG-02238 and the Huck Institute of Life Sciences of Penn State University, with additional support for L.E. from NHGRI grant HG-02325 and for R.C.H. from National Institute of Diabetes and Digestive and Kidney Diseases grant RO1 DK-27635; K.M.R., M.D., and W.J.K by NHGRI grant 1P41HG-02371; D.H. by NHGRI grant 1P41HG-02371 and the Howard Hughes Medical Institute.

REFERENCES

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., and Sherlock G. 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25:** 25.

Bailey J.A., Gu Z., Clark R.A., Reinert K., Samonte R.V., Schwartz S., Adams M.D., Myers E.W., Li P.W., and Eichler E.E. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003.

Bouhassira E., Westerman K., and Leboulch P. 1997. Transcriptional behavior of LCR enhancer elements integrated at the same chromosomal locus by recombinase-mediated cassette exchange. *Blood* **90:** 3332.

Chiaromonte F., Yap V.B., and Miller W. 2002. Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.* **2002:** 115.

Chiaromonte F., Yang S., Elnitski L., Yap V., Miller W., and Hardison R.C. 2001. Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proc. Natl. Acad. Sci.* **98:** 14503.

Collins F.S., Green E.D., Guttmacher A.E., and Guyer M.S. 2003. A vision for the future of genomics research. *Nature* **422:** 835.

Elnitski L., Hardison R.C., Li J., Yang S., Kolbe D., Eswara P., O'Connor M.J., Schwartz S., Miller W., and Chiaromonte F. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13:** 64.

Feng Y.Q., Seibler J., Alami R., Eisen A., Westerman K.A., Leboulch P., Fiering S., and Bouhassira E.E. 1999. Site-spe-

cific chromosomal integration in mammalian cells: Highly efficient CRE recombinase-mediated cassette exchange. *J. Mol. Biol.* **292:** 779.

Forget B.G. 2001. Molecular genetics of the human globin genes. In *Disorders of hemoglobin: Genetics, pathophysiology, and clinical management* (ed. M.H. Steinberg et al.), p. 117. Cambridge University Press, Cambridge, United Kingdom.

Friend C., Scher W., Holland J.G., and Sato T. 1971. Hemoglobin synthesis in murine virus-induced leukemic cells in vitro: Stimulation of erythroid differentiation by dimethylsulfoxide. *Proc. Natl. Acad. Sci.* **68:** 378.

Giardine B.M., Elnitski L., Riemer C., Makalowska I., Schwartz S., Miller W., and Hardison R.C. 2003. GALA, a database for genomic sequence alignments and annotations. *Genome Res.* **13:** 732.

Hardison R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369.

———. 2001. Organization, evolution and regulation of the globin genes. In *Disorders of hemoglobin: Genetics, pathophysiology, and clinical management* (ed. M.H. Steinberg et al.), p. 95. Cambridge University Press, Cambridge, United Kingdom.

Hardison R., Slightom J.L., Gumucio D.L., Goodman M., Stojanovic N., and Miller W. 1997. Locus control regions of mammalian β-globin gene clusters: Combining phylogenetic analyses and experimental results to gain functional insights. *Gene* **205:** 73.

Hardison R.C., Roskin K.M., Yang S., Diekhans M., Kent W.J., Weber R., Elnitski L., Li J., O'Connor M., Kolbe D., Schwartz S., Furey T.S., Whelan S., Goldman N., Smit A., Miller W., Chiaromonte F., and Haussler D. 2003. Covariation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. *Genome Res.* **13:** 13.

Jordan I.K., Rogozin I.B., Glazko G.V., and Koonin E.V. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* **19:** 68.

Kent W.J., Baertsch R., Hinrichs A., Miller W., and Haussler D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100:** 11484.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., and Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217:** 624.

Korf I., Flicek P., Duan D., and Brent M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* (suppl. 1) *17:* S140.

Lander E.S. 1996. The new genomics: Global views of biology. *Science* **274:** 536.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., and Naylor J., et al. (International Human Genome Sequencing Consortium). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860.

Lercher M.J. and Hurst L.D. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18:** 337.

Li J. and Miller W. 2002. Significance of interspecies matches when evolutionary rate varies. *J. Comput. Biol.* **10:** 537.

Li Q., Peterson K., Fang X., and Stamatoyannopoulos G. 2002. Locus control regions. *Blood* **100:** 3077.

Li W.H., Gojobori T., and Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature* **292:** 237.

Matys V., Fricke E., Geffers R., Gossling E., Haubrock M., Hehl R., Hornischer K., Karas D., Kel A.E., Kel-Margoulis O.V., Kloos D.U., Land S., Lewicki-Potapov B., Michael H., Munch R., Reuter I., Rotert S., Saxel H., Scheer M., Thiele S., and Wingender E. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31:** 374.

Migliaccio A.R. and Papayannopoulou T. 2001. Erythropoiesis. In *Disorders of hemoglobin: Genetics, pathophysiology, and clinical management* (ed. M.H. Steinberg et al.), p. 52. Cambridge University Press, Cambridge, United Kingdom.

Molete J.M., Petrykowska H., Bouhassira E.E., Feng Y.Q., MIller W., and Hardison R.C. 2001. Sequences flanking hypersensitive sites of the beta-globin locus control region are required for synergistic enhancement. *Mol. Cell. Biol.* **21:** 2969.

Nekrutenko A., Makova K.D., and Li W.H. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: An empirical and simulation study. *Genome Res.* **12:** 198.

Onodera K., Takahashi S., Nishimura S., Ohta J., Motohashi H., Yomogida K., Hayashi N., Engel J., and Yamamoto M. 1997. GATA-1 transcription is controlled by distinct regulatory mechanisms during primitive and definitive erythropoiesis. *Proc. Natl. Acad. Sci.* **94:** 4487.

Pennacchio L.A. and Rubin E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2:** 100.

Pribnow D. 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci.* **72:** 784.

Schwartz S., Kent W.J., Smit A., Zhang Z., Baertsch R., Hardison R.C., Haussler D., and Miller W. 2003. Human-mouse alignments with *Blastz. Genome Res.* **13:** 103.

Smit A. and Green P. 1999. *RepeatMasker* at: http://ftp.genome.washington.edu/RM/RepeatMasker.html

Stamatoyannopoulos G. 2001. Molecular and cellular basis of hemoglobin switching. *Disorders of hemoglobin: Genetics, pathophysiology, and clinical management* (ed. M.H. Steinberg et al.), p. 131. Cambridge University Press, Cambridge, United Kingdom.

Su A., Cooke M., Ching K., Hakak Y., Walker J., Wiltshire T., Orth A., Vega R., Sapinoso L., Moqrich A., Patapoutian A., Hampton G., Schultz P., and Hogenesch J. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci.* **99:** 4465.

Surinya K.H., Cox T.C., and May B.K. 1998. Identification and characterization of a conserved erythroid-specific enhancer located in intron 8 of the human 5-aminolevulinate synthase 2 gene. *J. Biol. Chem.* **273:** 16798.

Thomas J.W., Touchman J.W., Blakesley R.W., Bouffard G.G., Beckstrom-Sternberg S.M., Margulies E.H., Blanchette M., Siepel A.C., Thomas P.J., McDowell J.C., Maskeri B., Hansen N.F., Schwartz M.S., Weber R.J., Kent W.J., Karolchik D., Bruen T.C., Bevan R., Cutler D.J., Schwartz S., Elnitski L., Idol J.R., Prasad A.B., Lee-Lin S.Q., and Maduro V.V., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., Antonarakis S.E., Attwood J., Baertsch R., Bailey J., Barlow K., Beck S., Berry E., Birren B., Bloom T., Bork P., Botcherby M., Bray N., Brent M.R., Brown D.G., and Brown S.D., et al. (Mouse Genome Sequencing Consortium). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520.

Weiss M.J., Yu C., and Orkin S.H. 1997. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol. Cell. Biol.* **17:** 1642.

Wiehe T., Gebauer-Jung S., Mitchell-Olds T., and Guigo R. 2001. SGP-1: Prediction and validation of homologous genes based on sequence alignments. *Genome Res.* **11:** 1574.

Wilson M.D., Riemer C., Martindale D.W., Schnupf P., Boright A.P., Cheung T.L., Hardy D.M., Schwartz S., Scherer S.W., Tsui L.C., Miller W., and Koop B.F. 2001. Comparative analysis of the gene-dense ACHE/TFR2 region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucleic Acids Res.* **29:** 1352.

Wolfe K.H., Sharp P.M., and Li W.H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283.